ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

Movie-Sent: A Sentiment Classification Using Naive bayes and Support Vector Machine Techniques

Ashish A. Bhalerao[1] , Sachin N. Deshmukh[2] , Sandip D. Mali[3]

Research Student, Dept. of Computer Science and IT, Dr. Babasaheb Ambedkar Marthwada University, Aurangabad, MS, India.[1]

Professor, Dept. of Computer Science and IT, Dr. Babasaheb Ambedkar Marthwada University, Aurangabad, MS, India.[2]

Research Student, Dept. of Computer Science and IT, Dr. Babasaheb Ambedkar Marthwada University, Aurangabad, MS, India.[3]

ABSTRACT: The most important key of sentiments are sentiment lexicons that are the words present in review dataset called as opinion words. These words are generally used to identify positive or negative sentiments. A list of positive, negative and phrases is called as a sentiment lexicon or opinion lexicon. These type of sentiment lexicons and phrases are useful to identify sentiment. The input for sentiment analysis is customer reviews. There are huge number of reviews are available online. Sentiment analysis focus on these to generate differentiate between positive and negative. In this paper we focus on two methods are used in sentiment analysis such as Naïve Bayes and SVM. The results generated by this system are summarized and helpful for the user decision making. In this internet world, from the common man to a businessman; everyone is dependent on the Web. The opinions expressed on the web helps the users to determine which product or movie is good for them and it helps the businessman to determine what the customer thinks about their products. So, it is necessary to mine this large number of reviews and classify them, so it is helpful for them to read and take decisions. This movie-sent system automatic generate the rating, there is no need to read the review completely and take decisions, Also compare this technique and make a huge work of sentiment analysis and mention the scope of sentiment analysis for more research work.

KEYWORDS: sentiment analysis, opinion mining, sentiment classification

I. INTRODUCTION

Sentiment analysis, also called opinion mining. There are also many different names and different tasks, such as sentiment analysis, opinion mining, opinion extraction, sentiment mining, subjectivity analysis, affect analysis, emotion analysis, review mining, etc. However, they are now all under the sentiment analysis or opinion mining. [1] The growth of social media such as reviews, forum discussions, blogs, micro-blogs, Twitter, comments, and postings in social network sites on the Web, individuals and organizations are increasingly using the content in these media for decision making. Presently, if one wants to buy a consumer product, there is no need to asking to friends and family about opinions because there are many user reviews and discussions in public forums on the Web about the product. For an organization, it may no longer be necessary to conduct surveys, opinion polls, and focus groups in order to gather public opinions because there is a large quantity of such Information publicly available.[2] Each site typically contains a huge volume of opinion text that is not always easily translated in long blogs and forum postings. The average human reader will have difficulty identifying relevant sites and extracting and summarizing the opinions in them. Automated sentiment analyses systems are thus needed during an online shopping large number of reviews/opinions are given by the users that reflect whether the product is good or bad. These reviews need to be explored, analyse and organized for better decision making. Large number of user reviews or suggestions on everything is present on the web nowadays. Reviews may contain the reviews on different electronic products, books, user or critic reviews on movies etc. which

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

helps other users in their decision making. Reviews are increasing in a faster rate day by day because every person likes to give their opinion on the Web. Large numbers of reviews are available for a single product which makes difficult for a customer to read all the reviews and make a decision. Thus, mining this data, identifying the user opinions and classify them is an important task. Opinion Mining is a Natural Language Processing (NLP) and Information Extraction (IE) task that aims to obtain feelings of the writer expressed in positive or negative comments by analyzing a large number of documents. The main task of Sentiment analysis is to classify the documents and determine its polarity. Polarity is expressed as positive, negative or neutral.[3]

Sentiment classification is a task in text classification, in which there are two classes: positive and negative classes. Sentiment classification is conducted on the basis of sentiment rather than the topics as it is conducted in typical document classification. Sentiment classification techniques highly rely on sets of positive and negative lexicons to identify other indicative terms appearing in review opinions for classification. However, the positive and negative indicators vary across different kinds of consumer products. The sets of positive and negative indicators are not interchangeable but tailored for specific consumer products. Recent online consumer review systems usually provide user interfaces that separate the inputs of positive and negative comments rather than using only one input field in free text format. Sentiment classification becomes less important practically because the inputs of positive and negative opinions are submitted separately.[4]

There are three levels on which sentiment analysis can be performed such as, Document level classifies the whole document as positive, negative or neutral and commonly known as document-level sentiment classification. In this approach, it is assumed that the document contains an opinion on one main object expressed by the author of the document. Entire documents are classified as having an overall positive or negative polarity. Second, Sentence level classifies the sentences and identifies positive, negative or neutral opinion, commonly known as sentence-level sentiment classification. Neutral typically means no opinion. Third, Aspect level classifies sentences/documents as positive, negative or neutral based on the aspects of those sentences/documents commonly known as aspect-level sentiment classification. Aspect and feature level came into picture because document level mining and sentence level mining approaches can't find accurately what actually users likes and they does not like. It looks for sentiments on features of products.[5]

In this paper, we compare the different machine learning techniques with movie review dataset. The result of these techniques is totally depending on which dataset is used during execution of techniques.

II. RELATED WORK

Minqing Hu and Bing Liu has conducted experiments using the customer reviews of five electronics products: 2 digital cameras, 1 DVD player, 1 mp3 player, and 1 cellular phone. The reviews were collected from Amazon.com and C|net.com. Products in these sites have a large number of reviews. Each of the reviews includes a text review and a title. Additional information available but not used in this project includes date, time, author name and location (for Amazon reviews), and ratings. For each product, firstly crawled and downloaded the first 100 reviews. These review documents were then cleaned to remove HTML tags.[1] After that, NLProcessor [15] is used to generate partof-speech tags. The objective of this paper is to provide a feature-based summary of a large number of customer reviews of a product sold online. This experimental result indicates that the proposed techniques are very promising in performing their tasks. The average recall of opinion sentence extraction is nearly 70%. The average precision of opinion sentence extraction is 64%. WordNet is used as a dictionary to determine the opinion words and their synonyms and antonyms. Large numbers of movie reviews are collected from different-different websites. Movie reviews contain the user and critic reviews, there are various websites available on the web which contain movie. Final results are presented in graphical charts. To compute how well the system classifies each document as compared to human decision, all the documents were manually classified and the corresponding opinion was determined. The results were then compared with the result of the system. Same reviews were also applied to the other system named as "AIRC Sentiment Analyzer" available online[14]. Finally the results of the two systems were compared and the results have shown that the results of the document based Sentiment orientation system are better than that of AIRC Sentiment Analyzer. [2] The use of 'Adverb+Verb' combine with 'Adverb+Adjective' combine for document-level sentiment classification of a review. Though, the aspect-level sentiment profile produces a more focused and accurate sentiment summary of a particular movie and is more useful for the users. The results demonstrate that adding the sentiment score of 'Adverb + Verb' combines to the commonly used 'Adverb + Adjective' combine further improves the accuracy of sentiment classification. The best weight age factor for verb scores obtained through multiple experimental runs is 30%. We have

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

collected 10 reviews each for 100 Hindi movies from the popular movie review database website www.imdb.com.They have labelled all these reviews manually to evaluate performance of our algorithmic formulations. Out of 1000 movie reviews collected, 760 are labelled positive and 240 are labeled as negative reviews. [3]
Peter Turney[4] introduces a simple unsupervised learning algorithm for rating a review as thumbs up or down. The algorithm has three steps: (1) extract phrases containing adjectives or adverbs, (2) estimate the semantic orientation of each phrase, and (3) classify the review based on the average semantic orientation of the phrases. In experiments with 410 reviews from Epinions, the algorithm attains an average accuracy of 74%. It appears that movie reviews are difficult to classify, because the whole is not necessarily the sum of the parts; thus the accuracy on movie reviews is about 66%. On the other hand, for banks and automobiles, it seems that the whole is the sum of the parts, and the accuracy is 80% to 84%. Travel reviews are an intermediate case. Bo Pang and Lee[5] examine the effectiveness of applying machine learning techniques to the sentiment classification problem. A challenging aspect of this problem that seems to distinguish it from traditional topic-based classification is that while topics are often identifiable by keywords alone, sentiment can be expressed in a more subtle manner. For example, the sentence "How could anyone sit through this movie?" contains no single word that is obviously negative. Thus, sentiment seems to require more understanding than the usual topic-based classification. So, apart from presenting results obtained via machine learning techniques, author also analyzes the problem to gain a better understanding of how difficult it is. Ding, Liu [6] has focus on task to decide whether the comments are positive or negative That is, given a set of product features of a product, he has accurately identify the semantic orientations of opinions expressed on each product feature by each reviewer. The method basically counts the number of positive and negative opinion words that are near the product feature in each review sentence. If there are more positive opinion words than negative opinion words, the final opinion on the feature is positive and otherwise negative. Dave, Lawrence [7] describes a tool for sifting through and synthesizing product reviews, automating the sort of work done by aggregation sites or clipping services. We begin by using structured reviews for testing and training, identifying appropriate features and scoring methods from information retrieval for determining whether reviews are positive or negative. These results perform as well as traditional machine learning methods. Then use the classifier to identify and classify review sentences from the web, where classification is more difficult. Min Kim [8] has address the challenge problem in sentiment analysis that is given a Topic (e.g.,"Should abortion be banned?") and a set of texts about the topic, find the Sentiments expressed about (claims about) the Topic (but not its supporting subtopics) in each text, and identify the people who hold each sentiment. To avoid the problem of differentiating between shades of sentiments, simplify the problem to: identify just expressions of positive, negative, or neutral sentiments, together with their holders. For these 100 sentences were selected from the DUC 2001 corpus with the topics "illegal alien", "term limits", "gun control", and "NAFTA". Two humans annotated the 100 sentences with three categories (positive, negative, and neutral. Harb[09] has proposed a new approach for automatically extracting positive and negative adjectives in the context of opinion mining. Experiments conducted on training sets (blogs vs. cinema reviews) showed that our approach was able to extract relevant adjectives for a specific domain. Liu, Jingjing and Stephanie Seneff[11] have discussed a parse-and-paraphrase concept to measure the degrees of sentiment for product reviews. In this paper, they suggest a method to retrieving or extracting noun phrases adverb, adjective, based on clause structure found by parsing sentences into a hierarchical description. He has suggested a robust method for modelling the participation of adverb and negative words to the score for sentiment analysis. In this application 45% progress obtained using participating retrieving or extracting aspect-based from restaurant reviews. Sentiment identification have been considered; though, the majority of previous work deliver binary polarities such as positive and negative, and the polarity of sentiment is just reversed at a time of negation is identified.

III. PROPOSED METHODOLOGY

There are three levels of sentiment analysis i.e., Document level, Sentence level and Aspect level. In this thesis we are going to discuss our work which is related with Document level sentiment analysis. It classifies the whole document as positive or negative and commonly known as document-level sentiment classification. In this chapter we study the Movie-Sent approach of document level sentiment analysis using dataset. In this experiment work we used movie review dataset. Today huge numbers of users are interested to check the movie rating on different website before watching the movie. They are aware about all these things. Maximum website upload the movie review with ratings; such as 7.2 out of 10 or four star out of five star. In this work we are mostly focus on auto generate the rating score of any movie. Our experiment is not only limited to specific movie dataset. With the help of this Movie-Sent system and

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

the dictionary of positive and negative words, we are able to determine the rating score of any movie. The process or methodology of Movie-sent system shows in following figure.

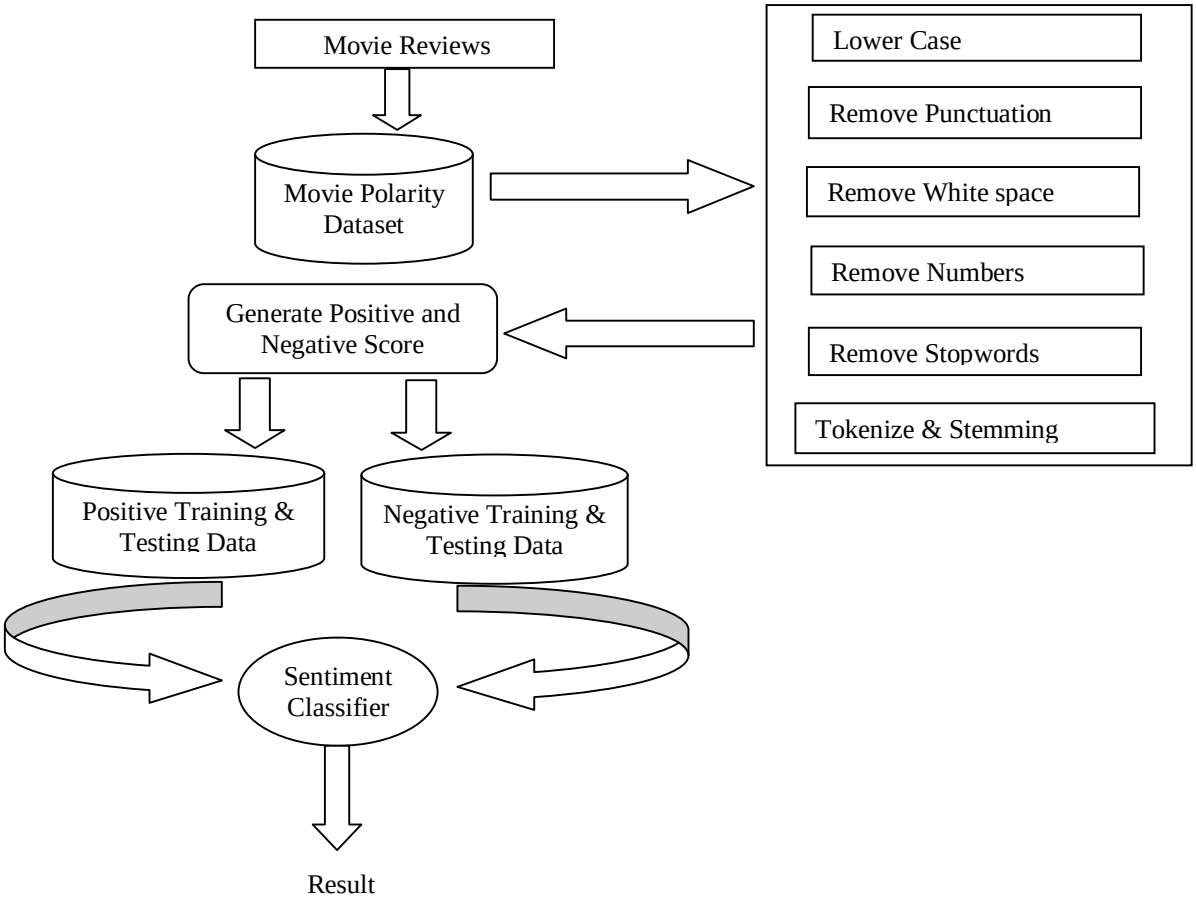


Figure 3.1: Architecture of a Movie-Sent system

The following are methods we use to implement the mining movie reviews using sentiment Analysis.

A. DATASET

Use dataset of Movies Reviews from IMDB, There are 1400 movie reviews in which 700 are positive and 700 are negative movie reviews.

B. DATA PRE-PROCESSING

Pre-processing on movie review dataset by using following techniques,

- Reading single or multiple files.
- Skip white spaces
- Remove Numbers
- Remove Punctuation
- Remove stop words
- Tokenization
- Stem Document
- Create Corpus

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

C. TO GENERATE THE SCORE OF MOVIE REVIEWS AS POSITIVE, NEGATIVE, NEUTRAL:

We generate a single score for each movie. It shows the final count of positive or negative words are present in this file. The document is positive if final score count is greater than 1 and the document is negative if final score count is lower than 1.

D. CLASSIFICATION OF SCORE OF USER REVIEWS.

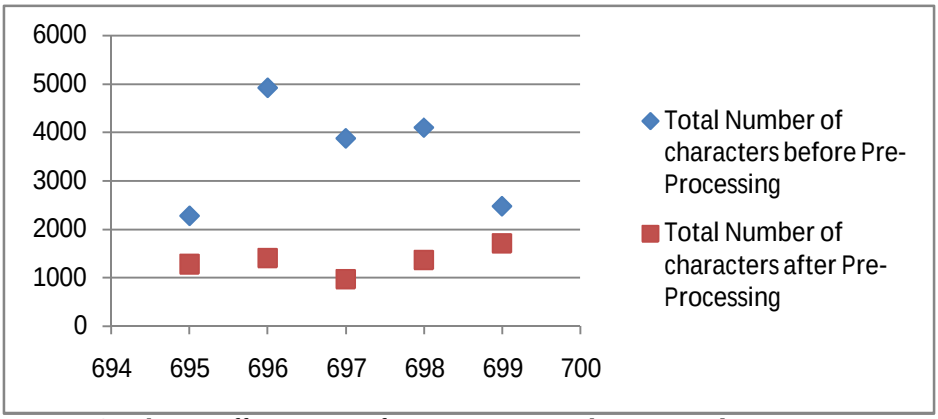
We apply various methods and algorithms such as Naïve bayes and SVM on these selected score**.**

E. RESULT

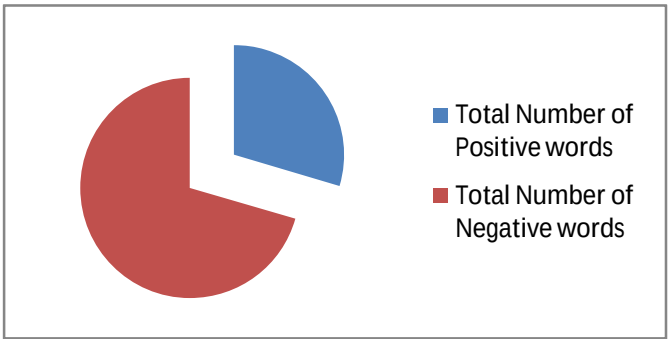
Result is shows that document is positive or negative and compare our system rating result with IMDB movie review rating website

IV. EXPERIMENTAL RESULTS WITH TABLES AND GRAPHS

We used movie review dataset available by Pang and Lee. There are total 14000 English movies review documents in which 700 are positive review and 700 are negative review. Basically this is IMDB movie review dataset. IMDB is a very popular website where user can put reviews and rating about movie. Any user can check the movie rating by accessing this website. This type of ratings is available in number such as (Rating 7.8 out of 10). Thus user can easily imagine the popularity of movie.



Graph 4.1: Effectiveness of Pre-processing techniques on dataset corpus.



Graph 4.2: Total 2003 Positive and 4776 Negative words dictionary used in Movie-Sent System

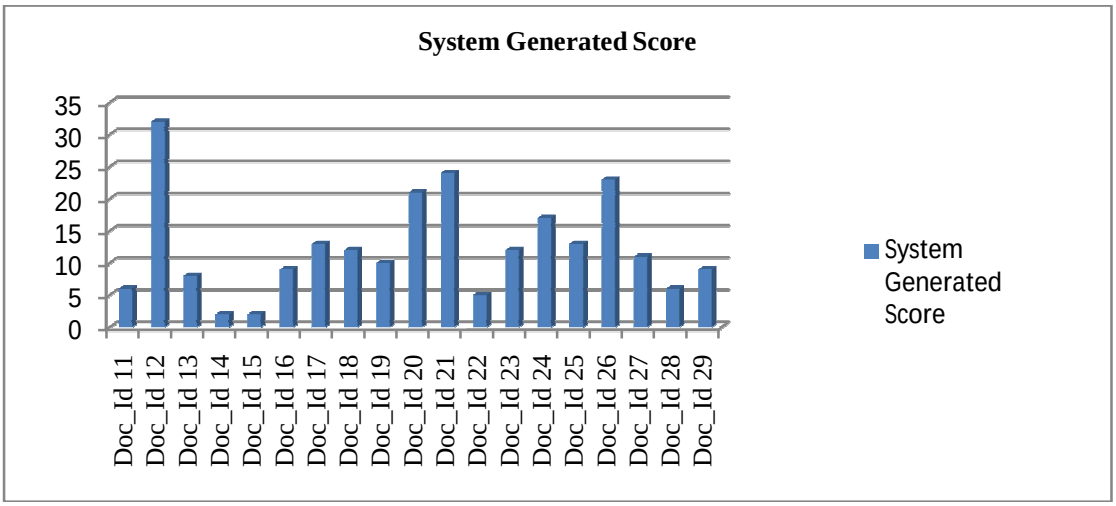
In the next step, we mentioned the final rating score by using score algorithm. The above mentioned steps are very important to generate the appropriate score of the movie review document.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

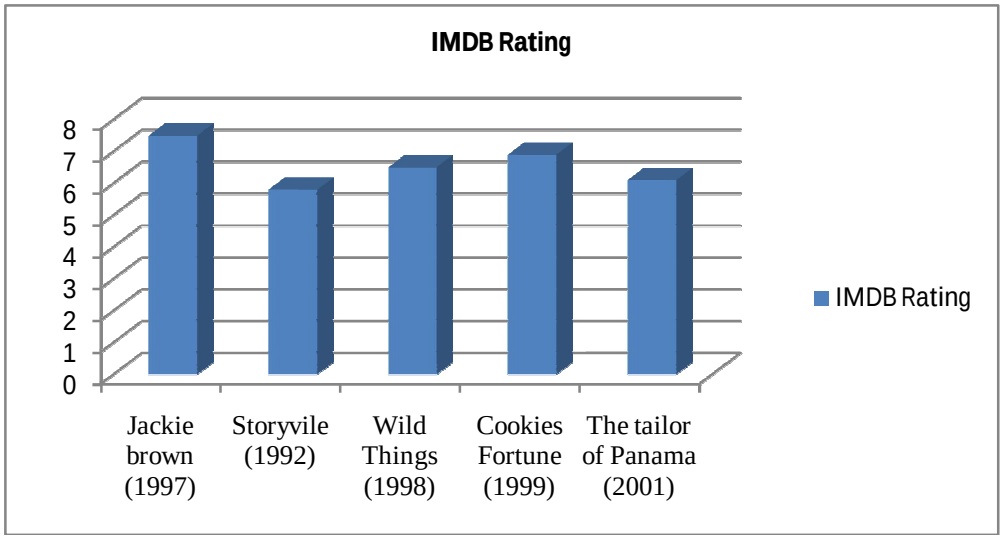


Graph 4.3: System Generated Score of Positive Movies Documents

Following table shows the highest rating obtained by IMDB 7.5 and for this same movie our system shows the highest rating i.e. 15. So with the help of our Movie-sent system user can imagine the movie is positive or negative with rating score.

Movie	IMDB Rating	Movie-Sent System Rating Score
Jackie brown (1997)	7.5	15
Storyvile (1992)	5.8	10
Wild Things (1998)	6.5	12
Cookies Fortune (1999)	6.9	14
The tailor of Panama (2001)	6.1	12

Table 4.1: Comparison of IMDB Rating and Movie-Sent System Rating Score



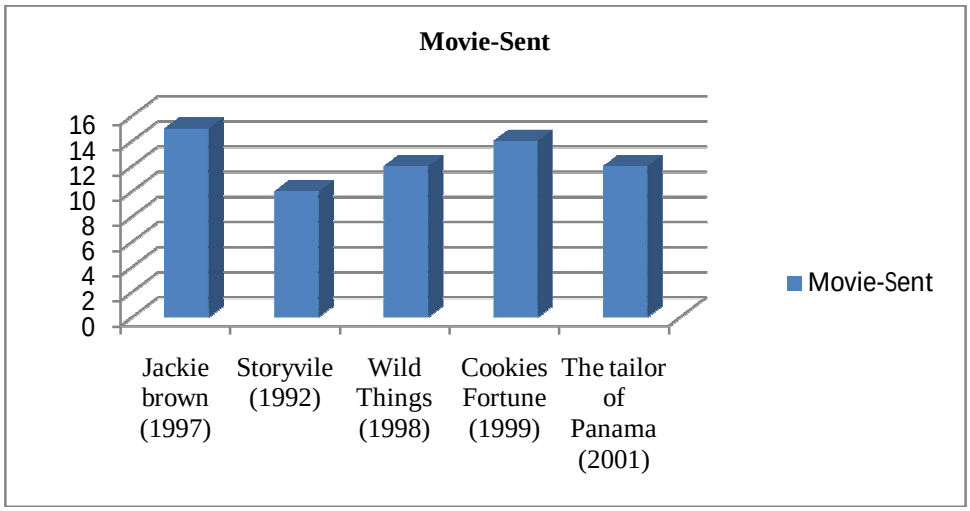
Graph 4.4: IMDB Review of Positive Rating movies

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

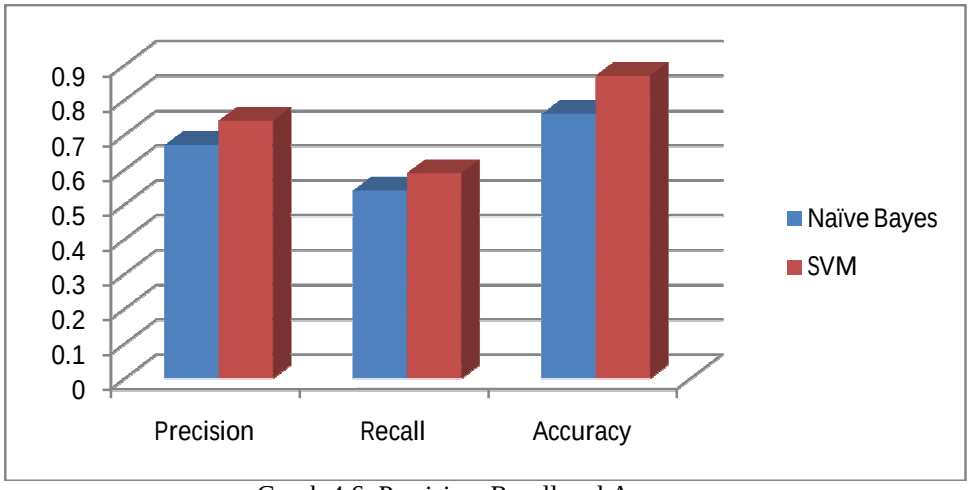
Vol. 4, Issue 11, November 2016



Graph 4.5: Movie-Sent System Score of Positive Rating movies

In the next step, we used two different machine learning techniques such as Naive bayes and Support Vector Machine (SVM) to determine the sentiment of document. For the classification purpose we divide our data into 70% for training purpose and 30% for testing purpose. The training data has 490 positive and 490 negative review documents and for testing purpose we use 30% of the data, which is 210 for positive testing data and 210 for negative testing data. There are total 980 training data and total 420 testing data.

We have calculated Precision, Recall and accuracy of naive bayes and SVM algorithm. In following table Naive bayes algorithm provides Precision 0.67, Recall 0.54 and Accuracy 0.76 and SVM provides Precision 0.74, Recall 0.59 and Accuracy 0.87.



Graph 4.6: Precision, Recall and Accuracy

V. CONCLUSION AND FUTURE WORK

IMDB is a popular online movie rating website. Various users are checking the rating before watching the movies. Our Movie-Sent system effectively shows the rating compare as IMDB website ratings. There is no need to read all reviews. This system completely works as automated rating system. There are various techniques available on different domain of dataset. We used Naive Bayes and Support vector machine algorithm. From this paper we conclude that, Support vector machine provides best result than Naive bayes on IMDB movie review dataset. We get better accuracy using Naive bayes which is good comparatively previous result which we seen in literature review. In future work,

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 11, November 2016

apply Movie-sent system on various other movie review dataset and cross domain dataset and compare the effectiveness of system.

VI. ACKNOWLEDGEMENT

We authors sincerely thanks to Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad for providing laboratory facility to complete this work.

REFERENCES

1. M. Hu and B. Liu, "Mining and summarizing customer reviews," 2004 presented the Proceedings of the tenth ACM.
2. Richa Sharma, Shweta Nigam and Rekha Jain, "Opinion mining of Movie Reviews at Document Level" International Journal on Information Theory (IJIT) Vol 3, No. 3 , July 2014
3. Singh, V.K ; Piryani, R ; Uddin, A., "Sentiment Analysis of movie reviews: A new feature based heuristic for aspect-level sentiment classification", IEEE, Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), 2013 International Multi-Conference on Jan 2013 ;
4. P. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews", Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics, pp. 417-424, Philadelphia, US, 2002
5. B. Pang, L. Lee & S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques", Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 79-86, Philadelphia, US, 2002.
6. Ding, Xiaowen, Bing Liu, and Philip S. Yu. A holistic lexicon-based approach to opinion mining. in Proceedings of the Conference on Web Search and Web Data Mining (WSDM-2008). 2008
7. K. Dave, S. Lawerence & D. Pennock, "Mining the Peanut Gallery-Opinion Extraction and Semantic Classification of Product Reviews", Proceedings of the 12th International World Wide Web Conference, pp. 519-528, 2003.
8. S.M. Kim & E. Hovy, "Determining sentiment of opinions", Proceedings of the COLING Conference, Geneva, 2004.
9. Ali Harb, Michel Planti_e, G_erard Dray, Mathieu Roche, Fran_cois Trousset, et al.. Web Opinion Mining: How to extract opinions from blogs?. CSTST'08: International Conference on Soft Computing as Transdisciplinary Science and Technology, pp.7, 2008.
10. Mejova, Yelena and Padmini Srinivasan. Exploring Feature Definition and Selection for Sentiment Classifiers. in Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM-2011). 2011.
11. Liu, Jingjing and Stephanie Seneff. Review sentiment scoring via a parseand-paraphrase paradigm. in Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP-2009).2009.
12. Gamon, Michael. Sentiment classification on customer feedback data:noisy data, large feature vectors, and the role of linguistic analysis. In Proceedings of International Conference on Computational Linguistics (COLING-2004). 2004.
13. www.imdb.com
14. http://airc-sentiment.org/
15. NLProcessor – Text Analysis Toolkit. 2000. http://www.infogistics.com/textanalysis.html
16. Wikipedia. Sentiment analysis. http://en.wikipedia.org/wiki/Sentiment_analysis.
17. http://www.searchenginepeople.com/blog/12-statistics-on-consumer-reviews. html#ixzz2DkFWhnBg
18. Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. Found. Trends Inf. Retr., 2(1-2):1–135, January 2008.